CS M148 Project 3 Report

Bonnie Liu
005300989
Discussion 1B
6/1/2021

Executive Summary

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. The total cost of stroke in the US was $103.5 billion according to 2016 US dollar values. $68.5 billion or 66% of total cost was accounted for by indirect cost from underemployment and premature death. Age groups 45-64 years accounted for the greatest stroke related direct cost.

In this project, I worked with the Stroke Prediction Dataset found on Kaggle, which contained information of 5,110 individuals pertaining whether or not they had a stroke, gender, age, bmi, etc. My goal in this project is to leverage the data provided to come up with a prediction model that can accurately predict if an individual is at risk for stroke.

During the process of exploratory data analysis, I found that I was dealing with an imbalanced classification problem since only about 5% of the individuals in the dataset had a stroke. To deal with this, I downsampled the majority class. Additionally, the bmi feature was the only column with null values, so I imputed them with the median bmi value.

For feature augmentation, I explored how age, bmi, and average glucose level were all related to one another. For the binary categorical features like ever_married or Residence_type, I replaced the options with 0 and 1. For the categorical features with more than 2 categories like gender, work_type and smoking_status, I decided to one-hot encode them since the dimensionality of our dataset would not explode even after doing so. I ended up with 22 dimensions.

I then proceeded to try out a couple of different models and compared their performances. Although I reported the accuracy, precision, recall, F1 score, and confusion matrix for each model, I decided that the best metric for comparing model performance is F1 score since it takes into account both recall and precision.

The first method I tried was logistic regression on my augmented dataset. Then I used principal component analysis to reduce the dimensionality of my data from 22 features to 6 features. Doing so actually improved the performance of my logistic regression model.

Next, I tried two different ensemble methods using Decision Trees: Random Forest and bagging. Both performed better than a singular decision tree since a model based on a single decision tree tends to overfit the model. By taking the average over multiple decision trees or splitting the testing data into the different “bags”, we are able to prevent overfitting and improve our model performance.

Then I tried neural networks using sklearn's MLP Classifier. Initially, I kept getting the Convergence Warning, which means that optimization wasn't reached under the maximum number of iterations set. I continued increasing the parameter max_iter until it hit 2400 and consistently ran to completion. I also entertained different activation methods like sigmoid and tanh but ultimately stuck with ReLU because according to the sklearn documentation, ReLU works well for smaller datasets like ours. I achieved mediocre results with the MLP Classifier, but I wanted to see which model truly excelled by using K-fold cross validation.

I used 16 folds for my cross validation and did not want to increase the number of folds even more since the neural net took a significant amount of time to run to completion. Upon reporting the mean accuracy across folds, I found that logistic regression using PCA had the highest mean accuracy across folds with the ensemble method using bagging coming in close second.

Introduction

Background

According to the Center for Disease Control, stroke is a leading cause of death and a major cause of disability in the United States. Every year, about 795,000 people suffer from a stroke. 1 in every 6 deaths from cardiovascular disease was due to stroke in 2018.

Risk

If an individual has suffered from a stroke before, they are more likely to get it again. Some other conditions that increase the risk for stroke include high blood pressure, high cholesterol, heart disease, diabetes, sickle cell disease. There are also behaviors that increase the risk for stroke, including unhealthy diet, physical inactivity, obesity, too much alcohol, and tobacco usage. Other characteristics that increase the risk for stroke include whether someone else in their family has had a stroke before, age, sex, and race or ethnicity.

Attribute Information

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. ever_married: "No" or "Yes"
7. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. Residence_type: "Rural" or "Urban"
9. avg_glucose_level: average glucose level in blood
10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12. stroke: 1 if the patient had a stroke or 0 if not

Methodology

Part 0: Setting Up

I started by importing a bunch of Python modules and defining some functions that will come in handy later on in the process. Please note that these were the same as those provided in Project 2. Next, I read healthcare-dataset-stroke-data.csv.

Part 1: Basic Statistics

I ran the correlation matrix, and I found that ever_married and age are quite strongly correlated, which makes sense since people are more likely to be married after a certain age. Then I plotted scatter plots of the 3 numerical variables against each other, and I found that there seems to be a general positive correlation between age and bmi. By plotting some histograms, I found that our data is unbalanced. Just under 5% of the individuals in our dataset had stroke. This is problematic because in the future, our model could just guess “no stroke” 100% of the time and still get the correct answer 95% of the time. To deal with this issue, I researched a little and found that I had two options:

- Upsample the minority class.
    - Advantage: no loss of information
    - Disadvantage: possibility of overfitting
- Downsample the majority class
    - Advantage: runtime improved
    - Disadvantage: loss of information in majority class

I decided to downsample since we had 249 data points in our minority class, which I felt was enough for building a model.

Part 2: Data Feature Extraction

Upon reading healthcare-dataset-stroke-data.csv, the info() function showed that the bmi column had 201 missing values. To deal with this, I entertained a couple of options:

- Drop all rows with missing bmi values. 201/5110 is about 4% of the total data. This isn’t too bad, but let’s see if we can do better.
- Drop the bmi feature. Based on what I know about strokes and bmi, I have a gut feeling that these are correlated, so I decided against this option.
- Replace null bmi values with an arbitrary value like 0. While this is a quick and easy solution to our problem, I ultimately decided against this one as well since bmi is a numerical feature. Thus, we’ll want to perform linear regression on this feature, and replacing the null values with 0 will mess up our results.
- Replace null bmi values with values predicted by linear regression. This is perhaps the most effective solution, but given the fact that our dataset is quite large with 5110 entries, this may take a long time and may be computationally expensive.
- Replace null bmi values with the median or the mean. The describe() function showed that bmi data is skewed to the right, so I decided to replace null bmi values with the median rather than the mean.

I ultimately decided to replace the null bmi values with the median because of the aforementioned reasons. I also standardized our numerical features so that no single numerical feature will have a disproportionately large effect on the model. I one-hot encoded gender, work_type, and smoking_status because those were all categorical variables with a fairly low

number of categories, so our dimensionality did not inflate drastically. Lastly, I augmented the following features:

- age_over_bmi: As noted in Part 1, there seemed to be a general positive correlation between age and bmi.
- age_over_glucose: I thought it would be interesting to examine how average glucose levels and age affected each other.
- bmi_over_glucose: I thought it would be interesting to see if bmi and average glucose levels were related in any way.

I ended up with 22 columns total in my dataset after feature extraction.

Part 3: Logistic Regression

I decided to split my pipelined data in order to test my models. I allocated 80% of the data for the training set and 20% of the data for the testing set. For logistic regression, I decided to use the liblinear solver because it's good for smaller datasets. I also tried the default solver for logistic regression, which is lbfgs, and indeed, logistic regression using the liblinear solver performed better on our dataset. To analyze data, I printed out the following:

- Accuracy: the fraction of predictions our model got correct
- Precision: false positive rate
- Recall: false negative rate
- F1 score: a balance of precision and recall

I decided to go with F1 score as my main indicator of whether or not a model was good since it's the most generalized measure. A higher F1 score indicates better model performance.

Part 4: Principal Component Analysis

As mentioned above, my model had 22 columns, which is a lot, and by using principal component analysis, we can reduce the dimensionality in hopes of preventing overfitting. I played around with the parameter n_components (number of components) with values ranging from 3 to 10, and upon comparing F1 scores, I decided that n_components=6 produced the most optimal results.

Part 5: Ensemble Method

Before jumping into implementing an ensemble method, I wanted to see what a single decision tree looks like on our data. Then I proceeded to do a random forest with 10 decision trees. (I tried various values for the number of decision trees and settled at 10 because that was approximately where performance started to plateau.) Finally, I moved on to bagging by using the Bagging Classifier with the following parameters: Decision Tree Classifier, max_samples=0.3, and n_estimators=20. This meant that we would have 20 decision trees and that each "bag" could have a maximum of 30% of our training dataset. I came to these parameter values after a lot of experimenting (trial and error).

Part 6: Neural Net

I decided to use sklearn's MLP Classifier since it was the simplest to implement, but other (perhaps better) neural networks include pytorch and keras. I played around with the parameters of the MLP classifier. I kept getting ConvergenceWarning, which meant that the

optimization didn't converge under the current number of iterations, so I kept increasing max_iter until it hit 2000, which is pretty large compared to the default of 200. As a result of the large number of iterations my model had to go through, it took a while for my model to converge and my neural net to finish running. I did a little bit of research (see resources section below), and I found that ReLU converges faster than other activation functions like sigmoid and tanh, so I decided to stick with ReLU, which is the default.

Part 7: K-Fold Cross Validation

I ran K-fold cross validation on the logistic regression using PCA, the ensemble method using bagging, and the neural net using the MLP Classifier. These 3 were my best-performing models from this project, so I wanted to see how they would prefer over many iterations. I set shuffle to True and random_state to the same for all 3. For the number of folds, I set it to 16 since there were 498 rows in my data, so setting it to 16 would mean that each fold had about 31 data points. I was unwilling to create more folds because the neural net took a long time to run with 16 folds.

Part 8: Reporting Highest-Performing Model

Using the results from my K-fold cross validation, I printed out the mean accuracy across folds for each of the 3 models. I picked the model with the highest mean accuracy across folds as my best (i.e. highest-performing) model.

Results

Part 0: Setting Up

```
In [445]:  1 data.head()
```

Out[445]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

```
In [446]:  1 data.describe()
```

Out[446]:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

```
In [447]:  1 data.info()
```

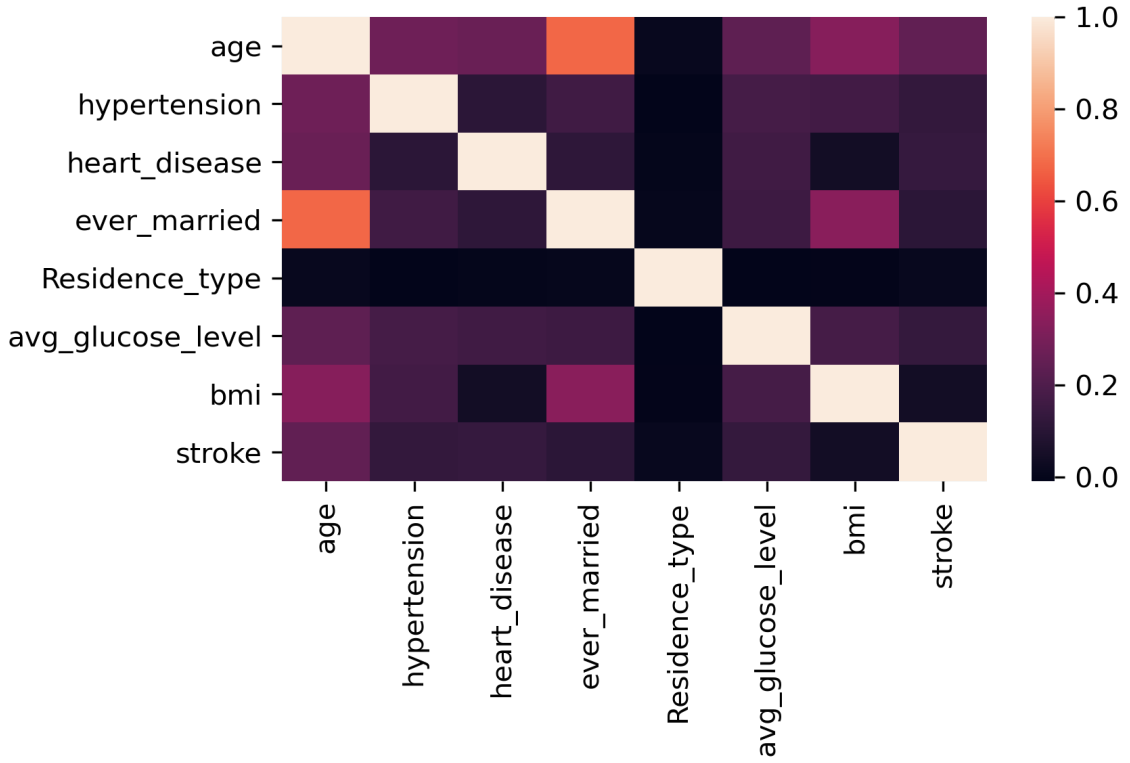
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

From the above basic Python data analysis functions, I was able to find that bmi was the only feature with null values. Additionally, I was able to conclude that age, avg_glucose_level, and bmi were the only numerical features whereas everything else was categorical.

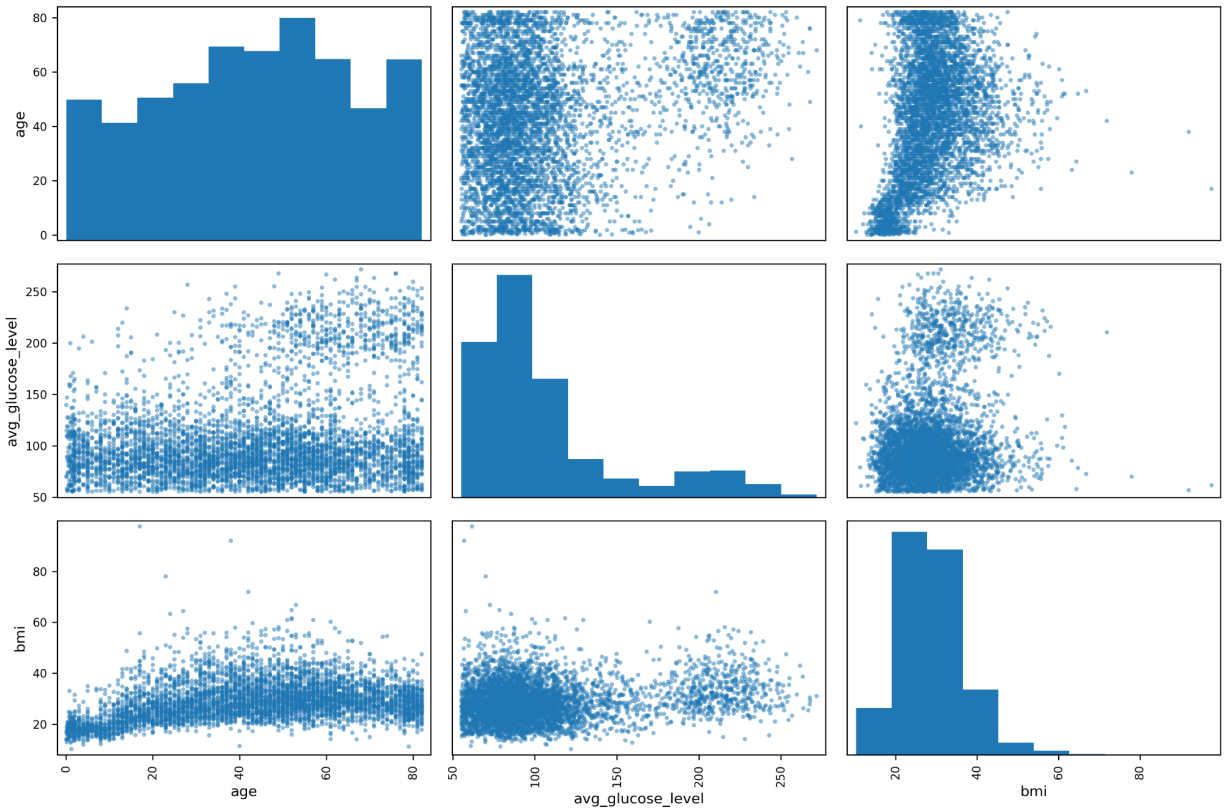
Part 1: Basic Statistics

In order to clean up the data, I used Label Encoder to change our binary categorical variables (i.e. stroke, ever_married, and Residence_type) into 0's and 1's. I also dropped the id column. This was my correlation matrix and heatmap afterwards:

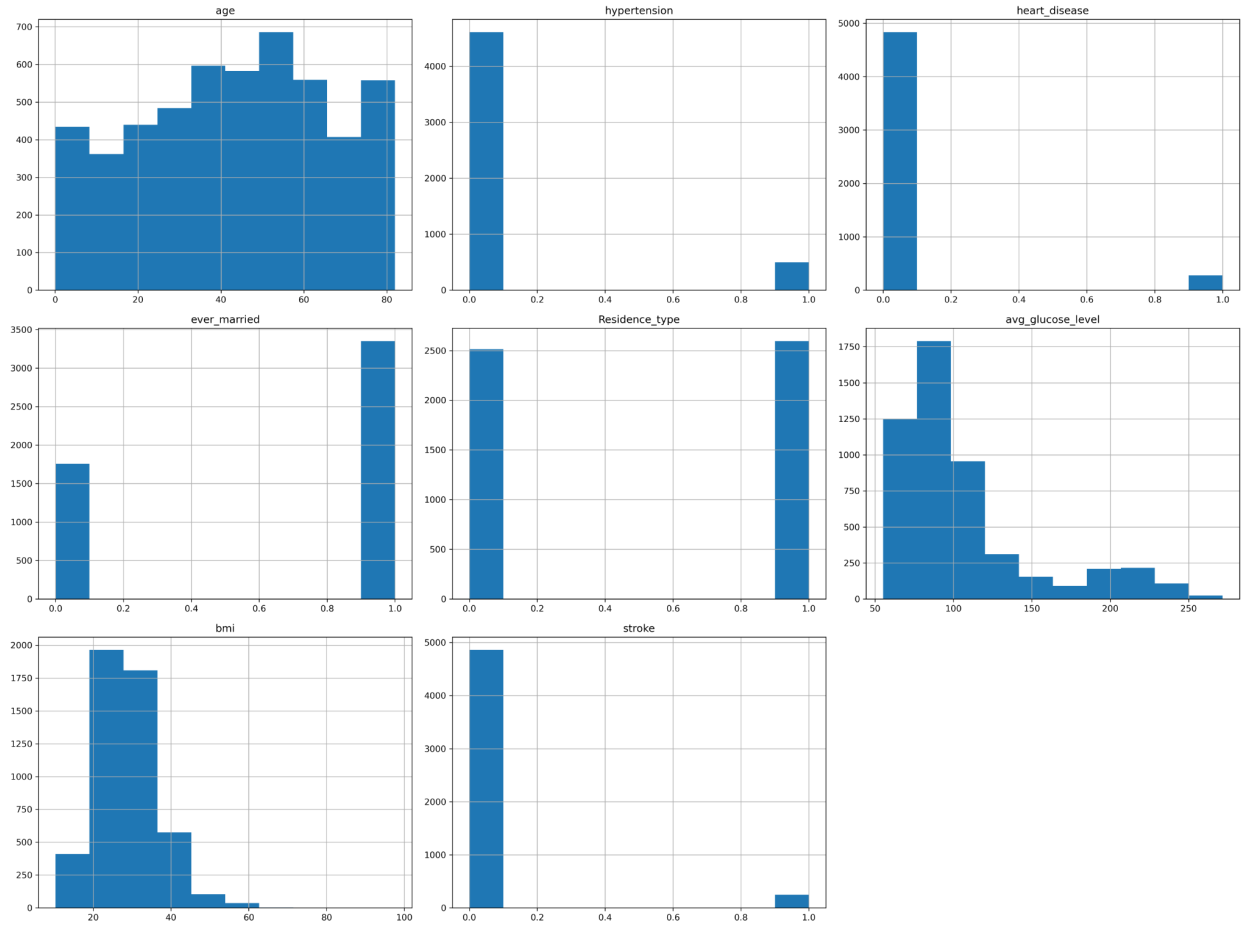
```
age                0.245257
hypertension       0.127904
heart_disease      0.134914
ever_married       0.108340
Residence_type     0.015458
avg_glucose_level  0.131945
bmi                0.042374
stroke             1.000000
Name: stroke, dtype: float64
```



As seen above, there seems to be a positive correlation between stroke and age, hypertension, heart_disease, avg_glucose_level, and marital status. I wanted to dive into the numerical features of the dataset and see if any of them were related, so I made scatterplots of age, avg_glucose_level, and bmi.



There seemed to be a positive correlation between age and bmi as well. It looks like perhaps a logarithmic relationship might exist between bmi and age. I proceeded by plotting histograms for each feature so that I could gain a better understanding of each feature individually.



As seen from above, bmi, age, and avg_glucose_level are gradients while hypertension, heart_disease, ever_married, Residence_type, and stroke are binary. Since stroke is our response variable, we know that we are dealing with a classification problem here. I decided to examine stroke in more detail, and I found that in our dataset, 249 individuals had a stroke while 4861 individuals did not. This means that just under 5% of our dataset has had a stroke and that our dataset is largely imbalanced. To deal with this, I decided to downsample the majority class as mentioned in the Methodology section. After doing so, I had 249 individuals with strokes and 249 individuals without.

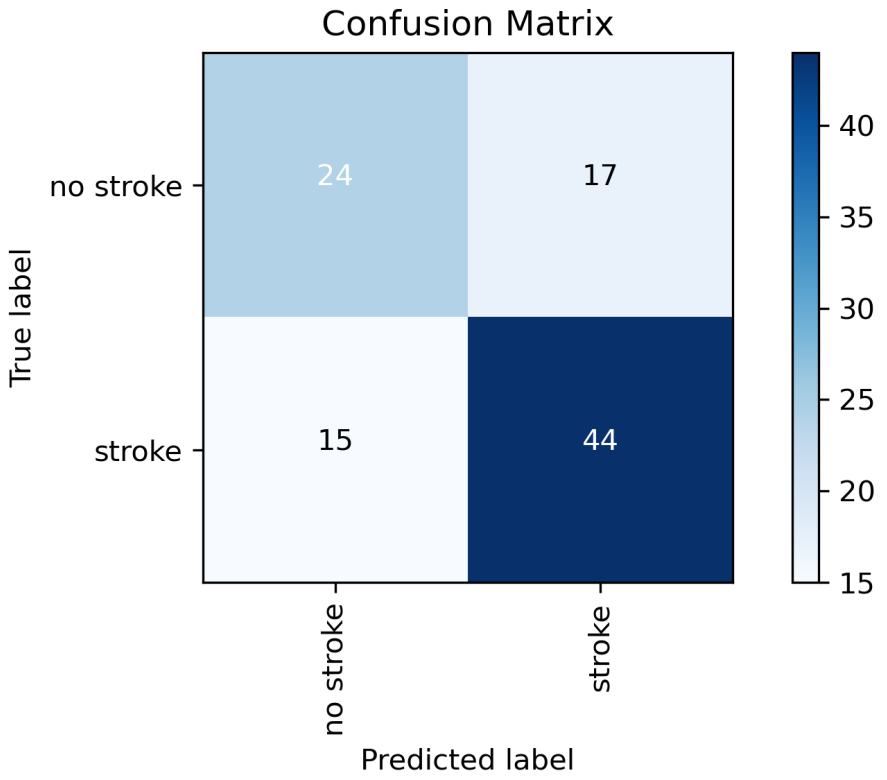
Part 2: Data Feature Extraction

After feature extraction using the pipeline, I ended up with a dataset with 498 rows and 22 columns.

Part 3: Logistic Regression

I split the dataset using an 80/20 split, so my training dataset had 398 data points while my testing dataset had 100.

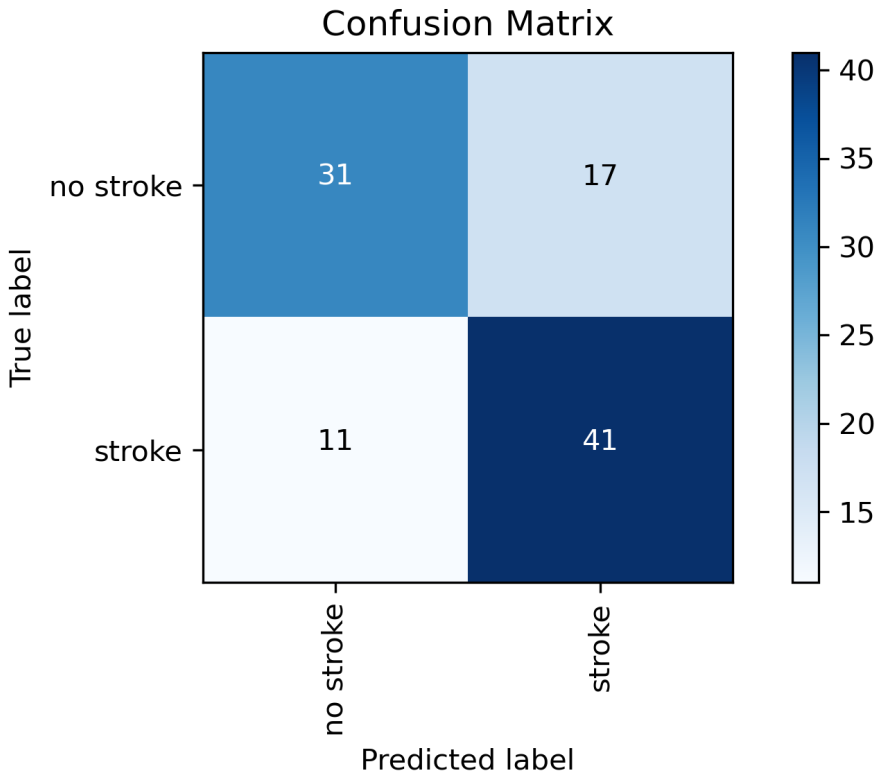
```
Accuracy:    0.680000
Precision:   0.721311
Recall:      0.745763
F1 Score:    0.733333
```



This isn't great considering the accuracy is pretty low, so let's see if we can do better after implementing Principal Component Analysis.

Part 4: Principal Component Analysis

```
Accuracy:    0.720000
Precision:   0.706897
Recall:      0.788462
F1 Score:    0.745455
```

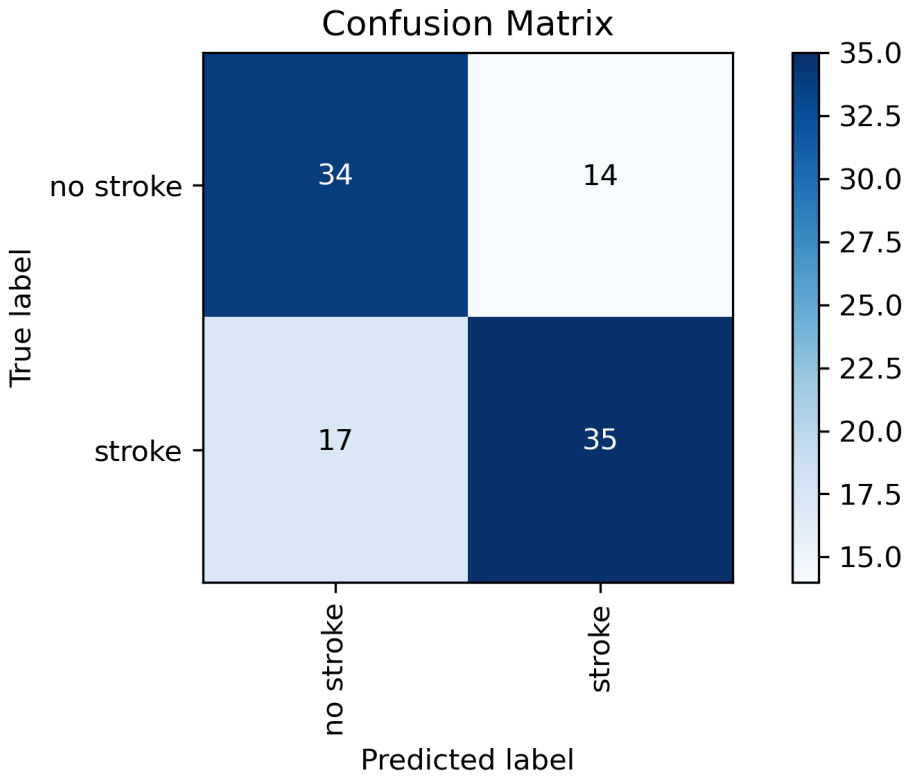


Indeed, our model improved when we reduced the dimensionality of the dataset from 22 columns to just 6. We got a better accuracy and F1 score as a result.

Part 5: Ensemble Methods

Before jumping into ensemble methods, I wanted to test out a single decision tree, and when I ran it, I got the following results:

```
Accuracy:    0.690000
Precision:   0.714286
Recall:      0.673077
F1 Score:    0.693069
```

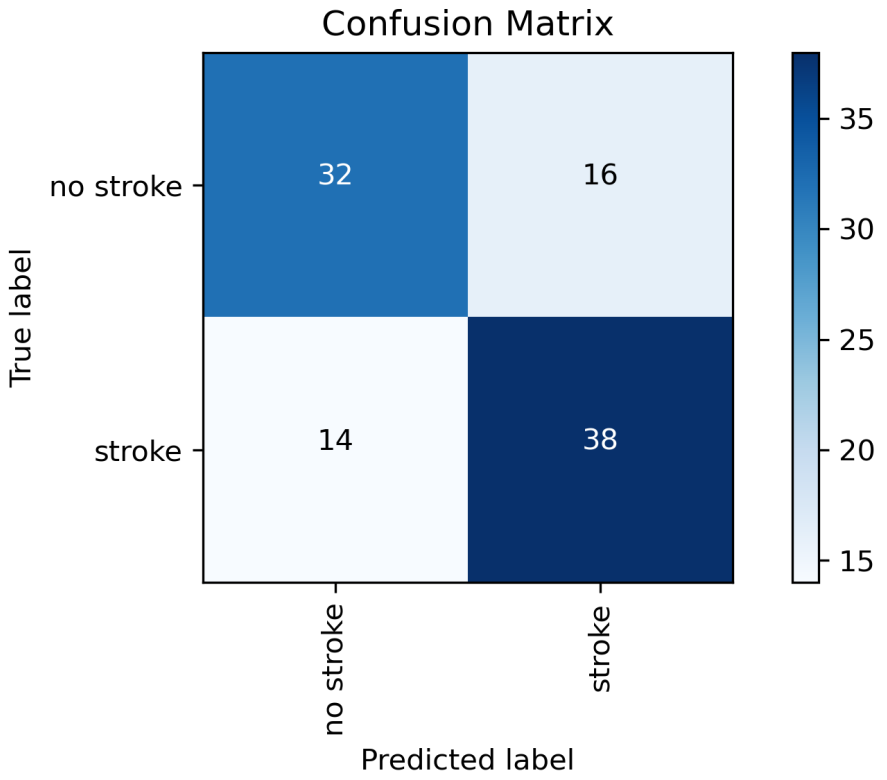


With a single decision tree, we can see that the model performance was not very good. I decided to dive into this situation a little further:

```
dt.score(new_X_train, new_y_train): 1.0
dt.score(new_X_test, new_y_test): 0.69
```

As we can see, it seems like our model is overfitting the training data. That's where ensemble methods come into play. Let's first try random forest, which is just a bunch of decision trees.

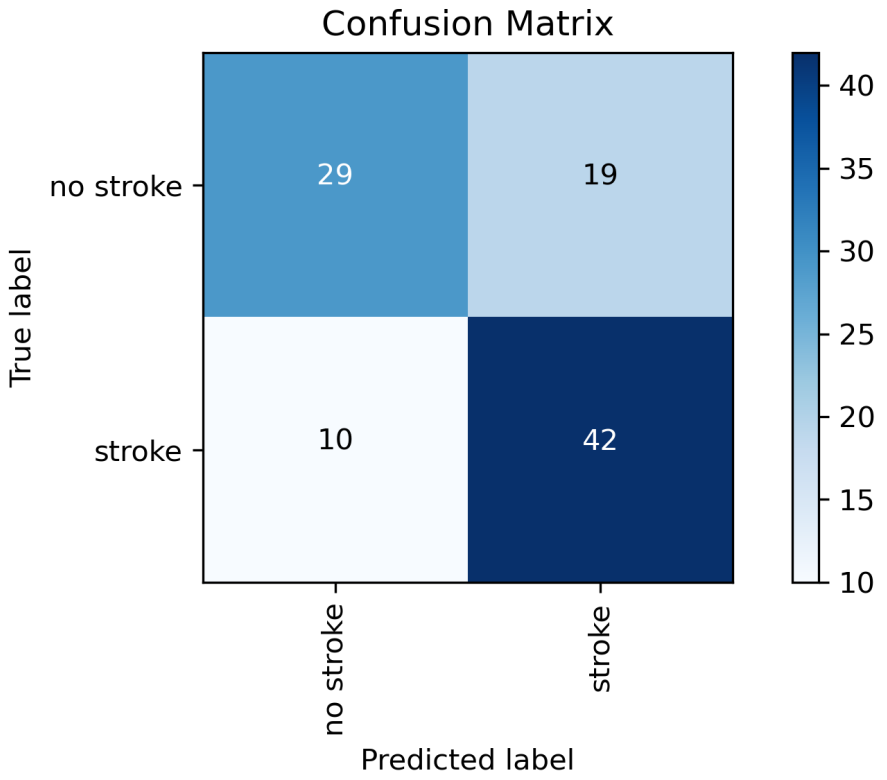
```
Accuracy:    0.700000
Precision:   0.703704
Recall:      0.730769
F1 Score:    0.716981
```




```
rf.score(new_X_train, new_y_train): 0.9899497487437185
rf.score(new_X_test, new_y_test): 0.7
```

As we can see, our random forest model does better than a single decision tree because it takes the average across all the decision trees. Our model is no longer overfitting the training data as much, which is consistent with our finding that this random forest performs better on the testing set than a single decision tree. Now let's try bagging to see if we can improve the performance even more.

```
Accuracy:    0.710000
Precision:   0.688525
Recall:      0.807692
F1 Score:    0.743363
```

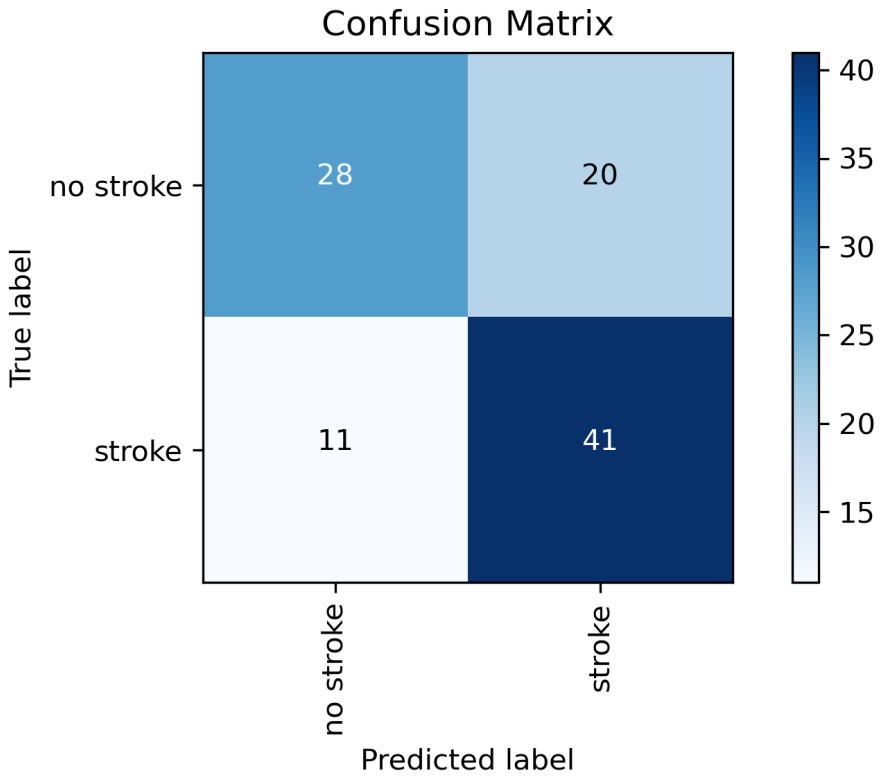



```
bg.score(new_X_train, new_y_train): 0.8592964824120602
bg.score(new_X_test, new_y_test): 0.71
```

Based on the above results, bagging drastically reduces overfitting and improves our model performance.

Part 6: Neural Net

```
Accuracy:    0.690000
Precision:   0.672131
Recall:      0.788462
F1 Score:    0.725664
```



Our neural network performed poorly compared to some of our earlier ensemble methods, but let's see if doing cross validation will help us determine which model is truly the best.

Part 7: K-Fold Cross Validation

For logistic regression using PCA, our mean accuracy across folds is 75.53%. For ensemble method using bagging, our mean accuracy across folds is 74.50%. For neural net using MLP Classifier, our mean accuracy across folds is 71.28%. As we can see here, logistic regression using PCA was our highest-performing model.

Discussion

Overall, I wasn't too satisfied with the results I obtained. I think my mean accuracy across folds definitely could have been higher, and if given more time to experiment with different models and tweak the parameters, I think I could have improved my accuracy. Some other methods I would like to explore for this dataset are k-nearest neighbors, support vector machines, and AdaBoost (ensemble method).

I would recommend the UCLA hospital to use my best-performing prediction model, which was the logistic regression using PCA since logistic regression is generally a good classification model for binary classification problems like these. I would advise the UCLA hospital to take the results from the model with a grain of salt because the model is far from perfect, and it would be best to have a doctor or an expert in the field additionally review the patient's information history.

In terms of next steps for analytic work, I think that while this dataset was quite comprehensive, more information could help this model improve even more. For example, based on the CDC's list of potential risk factors, it may be helpful to know whether an individual has suffered from a stroke before, their blood pressure, their cholesterol levels, whether or not they have diabetes, whether or not they have sickle cell disease. There is also behavioral information that could be included in this dataset, such as diet, physical activity levels, and alcohol levels. Characteristics like race or ethnicity could also be a big indicator of whether someone will get a stroke.

Conclusion

This project dove into various models including logistic regression, principal component analysis, decision trees, random forests, bagging, and neural networks to try to come up with an effective prediction model of whether an individual will have a stroke. Each model was trained on its respective training dataset, and each model's performance was evaluated using the testing dataset. F1 score was used as the main performance metric as it takes both precision and recall into account. This problem was an imbalanced classification problem, so I downsampled the majority class. After running K-fold cross validation on the various models I explored in this project, the model that performed best was the logistic regression model with PCA, which had approximately a 76% accuracy.

Resources

- Stroke | cdc.gov
- https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb
- https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5
- How to Handle Imbalanced Classes in Machine Learning (elitedatascience.com)
- Bagging and Random Forest for Imbalanced Classification (machinelearningmastery.com)
- Ensemble Learning, Bootstrap Aggregating (Bagging) and Boosting
- Scikit Learn Ensemble Learning, Bootstrap Aggregating (Bagging) and Boosting
- Bagging (Bootstrap Aggregation) - Overview, How It Works, Advantages (corporatefinanceinstitute.com)
- Chapter 10 Bagging | Hands-On Machine Learning with R (bradleyboehmke.github.io)
- sklearn.ensemble.BaggingClassifier — scikit-learn 0.24.2 documentation
- Multi Layer Perceptron | SKlearn | ipynb notebook example
- Activation Functions | Fundamentals Of Deep Learning (analyticsvidhya.com)